UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Rio Claro

Emílio Bergamim Júnior

Classificação semi-supervisionada com o modelo Potts utilizando uma aproximação de campo médio

Rio Claro - SP
2022

Universidade Estadual Paulista
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Emílio Bergamim Júnior

Classificação semi-supervisionada com o modelo Potts utilizando uma aproximação de campo médio

Trabalho de dissertação apresentado como requerimento à obtenção do título de mestre em Ciência da Computação pelo Programa Interunidades de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Fabricio Breve

Rio Claro - SP
2022

B493c

Bergamim Júnior, Emílio
Classificação semi-supervisionada com o modelo Potts utilizando uma aproximação de campo médio / Emílio Bergamim Júnior. -- Rio Claro, 2022
44 p. : il.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Fabricio Breve

1. Aprendizado semi-supervisionado. 2. Transdução. 3. Modelo Potts. 4. Aproximações de campo médio. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Geociências e Ciências Exatas, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Emílio Bergamim Júnior

Classificação semi-supervisionada com o modelo Potts utilizando uma aproximação de campo médio

Trabalho de dissertação apresentado como requerimento à obtenção do título de mestre em Ciência da Computação pelo Programa Interunidades de Pós-Graduação em Ciência da Computação

Comissão examinadora

Prof. Dr. Fabricio Breve(Orientador)

Prof. Dr. Marcos Gonçalves Quiles

Prof. Dr. Alexandre Luis Magalhães Levada

Conceito: Aprovado.

Rio Claro - SP, 17 de fevereiro de 2022

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

O aprendizado semi-supervisionado vem estabelecendo-se como uma área de pesquisa altamente relevante devido à sua capacidade de extrair informação de conjuntos que apresentam dados rotulados e não rotulados. Em particular, abordagens a esse problema que utilizam-se de um grafo para representar tal conjunto constituem um ramo de pesquisa intensamente ativo. Por outro lado, métodos de aprendizado de máquina baseados em modelos oriundos da mecânica estatística também já ocupam um lugar bem estabelecido na literatura, com especial destaque às aproximações de campo médio. Neste trabalho almeja-se realizar um estudo sobre o modelo Potts - originalmente proposto como uma descrição do magnetismo em sólidos - e uma aproximação de campo médio simples aplicados à classificação semi-supervisionada baseada em grafos. Assim, situa-se tal modelo entre outros métodos já aplicados ao problema e então são estudadas duas formas para ajuste do parâmetro β, sendo a inexistência de um procedimento como tal o principal empecilho à aplicabilidade prática deste modelo. Tais abordagens são então avaliadas individualmente e comparadas com algoritmos do estado da arte em cenários nos quais são variadas as construções do grafo de representação e a quantidade de pontos previamente classificados em diferentes conjuntos de dados. Os resultados obtidos indicam que o modelo Potts aproxima-se da acurácia de métodos do estado da arte conforme aumenta-se a quantidade de rótulos conhecidos e o número de arestas no grafo. Além disso, uma dessas abordagens utilizando o modelo Potts apresenta desempenho computacional consistentemente superior, sendo ao menos uma ordem de magnitude mais rápida que as demais quando avaliada em conjuntos comumente utilizados para *benchmarks* de classificação. **Palavras-chave**: Aprendizado semi-supervisionado. Transdução. Modelo Potts. Aproximações de campo médio.

Abstract

The field of semi-supervised learning has been establishing itself as a highly relevant research area due to its capacity of extracting information from sets that present labeled and unlabeled data. In particular, approaches to this problem that leverage on a graph to represent said sets consist of an active research area. On the other hand, machine learning methods based on models from physics hold a well-established place in the literature, with particular emphasis on mean-field methods. This work aims to study the Potts model - originally proposed as a description of certain types of magnetism in solids - and a mean-field approximation applied to the problem of graph-based semi-supervised classification. So, such model is situated alongside other methods already employed to the problem and then two approaches for tuning its β parameter are studied since the absence of a well defined approach to tune such parameter is the main limitation to its practical application. Such methods are evaluated and compared with state-of-the-art approaches in scenarios where the amount of labeled data available and the graphical representation are evaluated in different datasets. Obtained results point that the accuracy of the Potts model approaches the state-of-the-art as the amount of labeled data and the number of edges on the graph increase. In addition, one of the approaches with the Potts model presented a consistently superior computational performance in comparison to other methods, being at least one order of magnitude faster in sets commonly used for classification benchmarks. **Keywords**: Semi-supervised learning. Transduction. Potts model. Mean-field approximations.

Lista de ilustrações

Figura 1 – Diagrama do procedimento geral de classificação transdutiva baseada em grafos. 17
Figura 2 – Os três conjuntos artificais utilizados para avaliação do modelo Potts aplicado à classificação semi-supervisionada. Cores distintas indicam rótulos distintos. Em cima: conjunto de duas luas(esq.) e conjunto de três *clusters* (dir.); em baixo: conjunto de cinco gaussianas independentes. 28
Figura 3 – Resultados da acurácia, tempo de execução e energia interna dos algoritmos NMF utilizando quatro similaridades distintas. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados. 35
Figura 4 – Resultados da acurácia e tempo de execução em função do parâmetro k para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados. 36
Figura 5 – Resultados da acurácia e tempo de execução em função da taxa de instâncias previamente classificadas utilizando uma similaridade esparsamente conectada para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados. 37
Figura 6 – Resultados da acurácia e tempo de execução em função da taxa de instâncias previamente classificadas utilizando uma similaridade densamente conectada para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados. 38
Figura 7 – Resultados da acurácia e tempo de execução em função de r_l para conjuntos de alta dimensão utilizando a similaridade PNN. 39
Figura 8 – Resultados da acurácia e tempo de execução em função de r_l para conjuntos de alta dimensão utilizando a similaridade RBF. 40

Sumário

1 INTRODUÇÃO 10
1.1 Motivação 11
1.2 Objetivos 11
1.3 Organização do trabalho 12

2 APRENDIZADO SEMI-SUPERVISIONADO 13
2.1 Hipóteses do aprendizado semi-supervisionado 14
2.2 Classificação semi-supervisionada transdutiva 15
2.2.1 Transdução baseada em grafos 16
2.2.2 Construção de similaridades 19

3 MODELO POTTS 21
3.1 Inferência com o modelo Potts 21
3.2 Aproximações de campo médio 23
3.2.1 A aproximação variacional 24
3.3 Ajuste do parâmetro β 25

4 METODOLOGIA 27
4.1 Implementação 30

5 RESULTADOS E DISCUSSÃO 31
5.1 Conjuntos bidimensionais 31
5.2 Conjuntos de alta dimensão 33
5.3 Sobre o caso $\beta = 1$ 34

6 CONCLUSÃO 41

REFERÊNCIAS 42

1 Introdução

Os sucessivos aumentos da capacidade humana de aquisição de dados (MARR, 2018) levam à necessidade de métodos capazes de extrair informação relevante dos conjuntos adquiridos. Neste cenário, a pesquisa em torno do aprendizado de máquina é algo cada vez mais relevante de forma a desenvolver técnicas computacionais capazes de lidar com os problemas deste século.

Apesar dos campos de aplicação serem dos mais diversos, parte significativa dos métodos de aprendizado é focada em atividades de previsão (ou mais precisamente classificação) de determinados fenômenos. Para problemas desse tipo é geralmente necessário um *conjunto de treinamento* do qual é extraída a informação necessária à realização da tarefa de interesse. No entanto, muitos dos métodos utilizados nesse tipo de situação demandam um conjunto de treinamento suficientemente grande para obter êxito em sua tarefa (ENGELEN; HOOS, 2020).

Em situações na qual existe um grande número de dados no qual apenas uma pequena porção dos mesmos foi rotulada e deseja-se rotular o restante, o *aprendizado semi-supervisionado* vem se consolidando com uma abordagem sólida para o problema, em particular a classificação semi-supervisionada (ENGELEN; HOOS, 2020; CHONG et al., 2020). Nesse paradigma, o objetivo do aprendizado pode ser classificar todas as instâncias observadas (classificação transdutiva) ou obter uma regra de classificação para classificar qualquer instância apresentada ao computador (classificação indutiva).

Métodos como os descritos anteriormente devem boa parte de seu desempenho à representação escolhida para o conjunto de dados. Particularmente, este trabalho é um estudo do caso em que os dados são representados por um grafo: em oposição à descrição de um elemento por seus atributos no espaço original, os mesmos são identificados por suas relações com os demais elementos do conjunto. Conhecida por *aprendizado baseado em grafos*, essa abordagem consiste geralmente de uma dinâmica sobre a estrutura que descreve os dados (SUBRAMANYA; TALUKDAR, 2014) e converge para um estado que pode ser denominado *aprendizado*.

Mais especificamente, o presente trabalho é um estudo sobre o modelo Potts (WU, 1982) aplicado à classificação semi-supervisionada transdutiva baseada em grafos. Tal modelo foi proposto originalmente como uma descrição do magnetismo em sólidos e, assim como outros métodos (GABRIé; TRAMEL; KRZAKALA, 2015) e modelos (YANG; AHUJA, 2019) originados da física (em particular, da mecânica estatística) é recorrentemente estudado no campo de aprendizado de máquina.

1.1 Motivação

Trabalhos que empregam o modelo Potts ou similares em tarefas de aprendizado já são uma realidade há décadas (BLATT; WISEMAN; DOMANY, 1996; GETZ; SHENTAL; DOMANY, 2005; OTT; STOOP, 2006) e seu estudo continua um campo ativo de pesquisa. No caso semi-supervisionado, estudos recentes foram conduzidos nas frentes de agrupamento (SHI; LIU; ZHANG, 2018) e classificação (WANG et al., 2007; LI; WANG, 2016).

No caso de classificação semi-supervisionada (na qual existe um subconjunto dos dados cuja classificação é previamente conhecida e o objetivo é classificar o conjunto como um todo), os estudos previamente realizados utilizaram-se do modelo Ising, para o qual o modelo Potts é uma generalização. Além disso, tais trabalhos limitam-se a um estudo sobre o espaço de parâmetros do modelo e destacam a necessidade de desenvolver métodos para ajustar esses parâmetros de forma a extrair o máximo de suas possibilidades.

O aprendizado com o modelo Potts consiste de uma abordagem probabilística ao problema. No entanto, a distribuição de probabilidade que o define demanda um processo de normalização cuja solução exata é praticamente inviável. Nessas condições, duas abordagens foram historicamente utilizadas para lidar com o problema: métodos de Monte Carlo e aproximações de campo médio. Este trabalho será focado em abordagens do segundo tipo, que são determinísticas e, em alguns casos, exatas.

1.2 Objetivos

Frente às questões levantadas, o presente trabalho tem por objetivos:

- Situar o modelo Potts entre as demais abordagens de classificação semi-supervisionada. Em particular, através de uma revisão da literatura, argumenta-se que o mesmo consiste de um modelo de agrupamento com informação *a priori* acerca de a qual grupo determinados elementos pertencem.

- Utilizar uma aproximação de campo médio simples (mais especificamente, a aproximação variacional ou *Naive Mean Field Theory*) para contornar o problema de intratabilidade do modelo original.

- Analisar a dependência do algoritmo de classificação em relação ao parâmetro β, de forma a introduzir um procedimento de ajuste de parâmetros.

- Comparar os algoritmos propostos com as abordagens do estado da arte. Em particular, serão utilizados como *baselines Label Propagation* (ZHU; GHAHRAMANI, 2002) e *Local and Global Consistency* (ZHOU et al., 2004).

1.3 Organização do trabalho

O trabalho está organizado da seguinte forma: no Capítulo 1 discute-se o problema do aprendizado semi-supervisionado e o subproblema de transdução baseada em grafos. A seguir, no Capítulo 3 o modelo Potts é introduzido e situado entre os algoritmos discutidos no capítulo anterior. Na sequência é feita uma discussão sobre as aproximações de campo médio e os procedimentos de ajuste do parâmetro β. No Capítulo 4 apresenta-se a metodologia utilizada para avaliar os métodos de interesse e avaliar o desempenho destes quando comparados ao estado da arte na área. Em seguida, no Capítulo 5 os resultados são discutidos e seguidos pela conclusão do trabalho no Capítulo 6.

2 Aprendizado semi-supervisionado

As capacidades de efetuar operações lógico-aritméticas e armazenar informação em uma memória interna levantam questões acerca das capacidades de aprendizado dos computadores há tempos (MITCHELL, 1997). Afinal, a possibilidade de automatizar determinadas tarefas de forma que esta automação melhore com a experiência é tentadora.

Além disso, o aprendizado passa a ser necessário à realização de uma tarefa por um computador quando é inviável programá-lo especificamente para tal (ALPAYDIN, 2004). Considere o problema de classificação de imagens, no qual deseja-se determinar qual objeto está presente em uma imagem a partir de uma lista de objetos possíveis. Sem recorrer a métodos de aprendizado, seria necessária uma descrição analítica precisa dos objetos de interesse em termos da representação computacional adotada para a imagem. Em suma, o problema torna-se difícil devido à baixa compatibilidade entre as linguagens humana e de máquina.

Exposta a necessidade do aprendizado, é importante destacar o papel da experiência, definida como a quantidade de informação disponível no momento de realização da tarefa desejada (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). De forma a ser compreendido como um modelo de aprendizado é necessário que, à medida que mais informação torna-se disponível ao aprendiz (isto é, através da aquisição de experiência), o desempenho deste na execução da tarefa em questão seja aperfeiçoado (MITCHELL, 1997).

Não menos relevante ao aprendizado é a definição *a priori* do tipo de conhecimento que deseja-se aprender. Tal definição, que depende tanto dos objetivos da aplicação como da informação disponível (ALPAYDIN, 2004), leva a uma divisão do aprendizado de máquinas em diversas subáreas, sendo duas delas os aprendizados supervisionado e não-supervisionado.

De forma a discutir estes tipos de aprendizado, considere um conjunto de dados contendo informações de indivíduos saudáveis e portadores de uma doença específica. Para cada indivíduo, são fornecidos dados sobre sintomas relevantes ao diagnóstico da condição médica de interesse. Um algoritmo de *classificação* pode, através do conhecimento desses dados, obter uma regra para determinar a partir de um conjunto de sintomas se um novo indivíduo é ou não portador dessa condição.

Ainda considerando o mesmo conjunto, um algoritmo de *agrupamento* pode ser aplicado aos dados relativos a indivíduos portadores dessa condição de forma a separá-los em grupos cujos sintomas apresentem algum tipo de similaridade. Em comparação ao problema de classificação, na qual o objetivo é aprender uma regra de classificação, no agrupamento é aplicada uma regra que possibilita a divisão dos objetos em grupos

aproximadamente homogêneos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

Obter uma regra que permite atribuir uma classe a um objeto é uma tarefa de aprendizado supervisionado. De um conjunto contendo dados previamente classificados - o conjunto de treinamento - é extraída uma regra que permite classificar objetos ainda não observados (ALPAYDIN, 2004). Na separação de indivíduos doentes em subgrupos o objeto de interesse do aprendizado passa a ser uma possível estrutura organizacional dos dados, a qual deseja-se conhecer através do aprendizado. Esse tipo de atividade, também conhecida como estimação de densidade, é relativa ao aprendizado não-supervisionado (ALPAYDIN, 2004).

Um possível problema apresentado por algoritmos supervisionados é a dependência no tamanho do conjunto de treinamento. Apesar da alta quantidade de dados disponíveis, obter conjuntos grandes o suficiente e apropriadamente rotulados pode ser extremamente custoso (ZHU; GOLDBERG, 2009; CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Frente a esse desafio, o aprendizado semi-supervisionado emergiu como uma interessante área de pesquisa. Nessas abordagens, o aprendiz tem à sua disposição tanto dados rotulados como não-rotulados e ambos são utilizados para extrair informação do conjunto de interesse.

Voltando ao exemplo do conjunto de dados médicos, suponha uma situação na qual o diagnóstico de 10 pacientes seja conhecido e deseja-se diagnosticar outros 1000. Numa situação como essa, a obtenção de uma regra de classificação utilizando um método de classificação supervisionada pode não alcançar a generalização esperada para sua aplicação. É possível, no entanto, executar um algoritmo de agrupamento sobre o conjunto dos 1010 pacientes e em seguida aplicar uma regra de classificação o que, historicamente, foi uma das primeiras abordagens do aprendizado semi-supervisionado (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

2.1 Hipóteses do aprendizado semi-supervisionado

A discussão até aqui visou apresentar o contexto no qual o aprendizado semi-supervisionado pode vir a ser útil. No entanto, a ideia de aprender uma classificação valendo-se de dados não-rotulados pode parecer contraintuitiva. Recorda-se que, apesar de não-rotulados, tais dados ainda são representativos da estrutura geral do conjunto. Portanto, a hipótese fundamental para o apredizado semi-supervisionado é a existência de uma relação entre as características estruturais dos dados e seus rótulos (ENGELEN; HOOS, 2020).

As condições para a efetividade do aprendizado semi-supervisionado ainda são objeto de discussão. Argumenta-se, no entanto, que é suficiente a existência de uma estrutura de

grupos (tais como aquelas obtidas através de métodos não-supervisionados) na qual cada um destes representa uma classe (ENGELEN; HOOS, 2020). Esta é a chamada hipótese de agrupamento, sendo que algoritmos distintos podem estar associados a interpretações distintas da mesma. Pode-se então subdividir o aprendizado semi-supervisionado de forma a agrupar algoritmos que utilizam-se de uma mesma interpretação (ENGELEN; HOOS, 2020; CHAPELLE; SCHOLKOPF; ZIEN, 2006):

- **Hipótese de suavidade:** se dois pontos com atributos $\mathbf{x}_i$ e $\mathbf{x}_j$ são suficientemente próximos, seus rótulos y_i e y_j também são.
- **Separação de baixa densidade:** a superfície que separa classes distintas está situada em uma região cuja densidade de pontos é suficientemente pequena.
- **Hipótese das variedades:** embora os atributos estejam localizados em um espaço de (possivelmente) alta dimensão, pontos em uma mesma classe pertencem a uma estrutura de baixa dimensão (uma variedade).

Embora a apresentação até aqui tenha focado em exemplos de classificação semi-supervisionado, há também abordagens de agrupamento e regressão semi-supervisionados (ENGELEN; HOOS, 2020; CHAPELLE; SCHOLKOPF; ZIEN, 2006). O presente trabalho, porém, será focado em algoritmos de classificação. Mais especificamente, em um tipo de classificação conhecida como transdução.

2.2 Classificação semi-supervisionada transdutiva

Considere um conjunto de N pontos $D = \{(x_i, y_i)\}_{i=1}^{N}$, sendo $\{x_i\}_{i=1}^{N}$ uma representação dos atributos de cada objeto no conjunto e $\{y_i\}_{i=1}^{N}$ o rótulo associado a cada x_i. Um problema que permite a utilização de métodos de classificação semi-supervisionada é aquele no qual existem um subconjunto $D_l \subset D$ de dados rotulados e outro subconjunto $D_u = D \setminus D_l$ cujos rótulos são desconhecidos.

Como já discutido, o objetivo da classificação supervisionada é *indutivo*: deseja-se obter uma regra para atribuir a qualquer objeto ainda não observado um rótulo. Por vezes, este também é o objetivo da classificação semi-supervisionada, porém a formulação do problema permite que sejam aplicados métodos *transdutivos*, nos quais o objetivo é classificar apenas os pontos de D_u. Na taxonomia proposta por Engelen e Hoos (ENGELEN; HOOS, 2020), esta é a primeira distinção entre algoritmos semi-supervisionados.

Nesse sentido, métodos semi-supervisionados indutivos são mais similares a métodos de classificação supervisionada, visto que ambos possuem uma etapa de treinamento na qual a regra de classificação é aprendida. Já os métodos transdutivos estão mais próximos

de algoritmos de agrupamento uma vez que limitam-se à compreensão dos dados observados em D.

2.2.1 Transdução baseada em grafos

Métodos transdutivos consistem essencialmente de três etapas: a construção de um grafo G_D no qual cada vértice corresponde a um elemento de D, seguida pela construção de uma matriz $\mathbf{W}$ que atribui a cada aresta de G_D um peso associado à similaridade entre os vértices conectados pela aresta em questão e, por fim, a realização de uma inferência sobre estas duas estruturas.

Para satisfazer a hipótese de agrupamento, é necessário então que G_D e $\mathbf{W}$ sejam uma representação suficientemente descritiva das relações entre os objetos de interesse para possibilitar uma inferência adequada (ENGELEN; HOOS, 2020). A forma de utilização do grafo implica que tais métodos valem-se da hipótese de variedade, visto que os elementos passam a ser descritos por suas relações ao invés de seus atributos (ENGELEN; HOOS, 2020; CHAPELLE; SCHOLKOPF; ZIEN, 2006).

Outra característica essencial à transdução é a coerência com a informação contida em D_l. Esse requerimento dá origem, em diversos algoritmos, à noção de *propagação*: os rótulos previamente conhecidos são propagados para os elementos de D_u através de uma dinâmica em G_D (CHONG et al., 2020). Uma diagrama do procedimento geral de transdução é apresentado na Figura 1.

Nesse contexto, *Label Propagation* (LP) (ZHU; GHAHRAMANI, 2002) é um padrão na classificação transdutiva (ENGELEN; HOOS, 2020; CHONG et al., 2020) e tido como estado da arte por alguns autores (KARASUYAMA; MAMITSUKA, 2017; MA; XIAO; DENG, 2020). Tomando uma situação na qual cada elemento de D pode receber uma de q classificações distintas, parte-se de uma matriz de probabilidades $\mathbf{P}_{N\times q}$ e uma matriz de similaridade $\mathbf{W}_{N\times N}$ que atribui a cada aresta de G_D um peso. O algoritmo consiste de um método iterativo no qual $\mathbf{P}$ é atualizada de acordo com seus valores na iteração anterior ponderados por $\mathbf{W}$ enquanto as probabilidades associadas a elementos de D_l são mantidas fixadas, como detalhado no Algoritmo 2.2.1. Após convergência do procedimento, os pontos são classificados de acordo com seus rótulos mais prováveis:

$$y_i = arg\max_s P_{i,s}. \tag{2.1}$$

Outra abordagem também amplamente utilizada para a etapa de inferência é o algoritmo LGC (*Local and Global Consistency*) (ZHOU et al., 2004; CHONG et al., 2020; ENGELEN; HOOS, 2020) que corresponde ao problema de encontrar a matriz $\mathbf{F}_{N\times q}$ que minimiza a função

$$Q(\mathbf{F}) = \sum_{i<j} W_{i,j}||\mathbf{F}_i - \mathbf{F}_j|| + \frac{\alpha}{1-\alpha}\sum_i ||\mathbf{F}_i - \mathbf{Y}_i||^2 \tag{2.2}$$

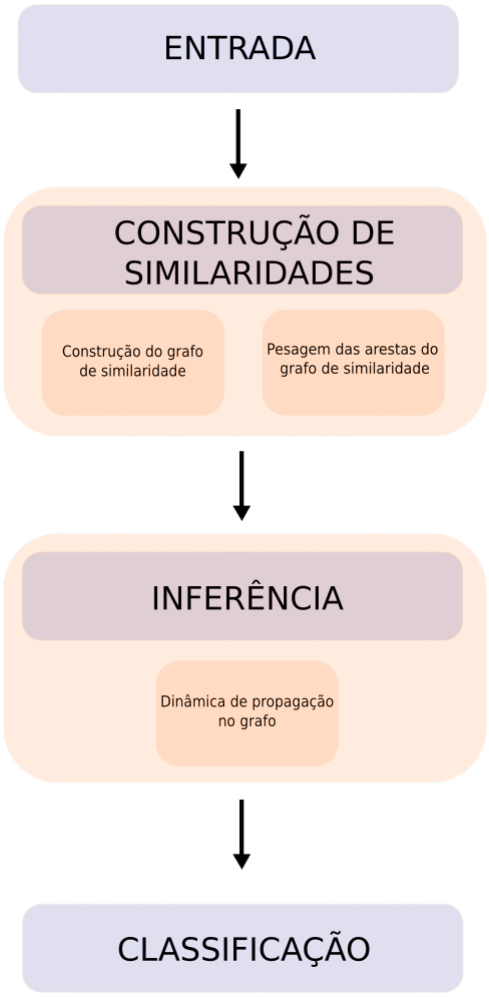

Figura 1 – Diagrama do procedimento geral de classificação transdutiva baseada em grafos.

onde $\mathbf{Y}_{N\times q}$ é a matriz de rótulos

$$Y_{i,k} = \begin{cases} 1 \text{ se } (x_i, y_i) \in D_l \text{ e } y_i = k, \\ 0, \text{ caso contrário,} \end{cases} \tag{2.3}$$

e $\alpha \in (0, 1)$ é o parâmetro do modelo.

A expressão (2.2) possui uma solução analítica que passa pela inversão de uma matriz (ZHOU et al., 2004; CHONG et al., 2020) e, portanto, tem complexidade $O(N^3)$. No entanto, o problema também pode ser resolvido iterativamente com complexidade $O(N^2)$, como exposto no Algoritmo 2.2.2. Ambas opções no entanto dependem do ajuste de α, que não é trivial(SOUSA; REZENDE; BATISTA, 2013). No caso de LGC, as classificações são obtidas de forma análoga ao algoritmo anterior:

$$y_i = arg\max_s F_{i,s}. \tag{2.4}$$

Algoritmo 2.2.1 Label Propagation

Entrada: $\mathbf{P}^{(0)}$, $\mathbf{W}$, D_l, ϵ, t_{max}
$t = 0$, $\delta = \epsilon$
while $t < t_{max}$ e $\delta \geq \epsilon$ **do**
  $\delta = 0$
  for $i = 1, ..., N$ **do**
    if $(x_i, y_i) \notin D_l$ **then**
      for $s = 1, ..., q$ **do**
        $P_{i,s}^{(t+1)} = \frac{\sum_{j \neq i} W_{i,j} P_{j,s}^{(t)}}{\sum_k \sum_{j \neq i} W_{i,j} P_{j,k}^{(t)}}$
      end for
    else
      for $s = 1, ..., q$ **do**
        if $s = y_i$ **then**
          $P_{i,s} = 1$
        else
          $P_{i,s} = 0$
        end if
      end for
    end if
    $\delta = \max\{\delta, \max_s |P_{i,s}^{(t+1)} - P_{i,s}^{(t)}|\}$
  end for
  $t = t + 1$
end while
Saída: $\mathbf{P}^{(t)}$

Algoritmo 2.2.2 LGC Iterativo

Entrada: $\mathbf{P}^{(0)}$, $\mathbf{W}$, $\mathbf{Y}$, α ϵ, t_{max}
$t = 0$, $\delta = \epsilon$, $\alpha = (1 + \mu)^{-1}$
while $t < t_{max}$ e $\delta \geq \epsilon$ **do**
  $\delta = 0$
  for $i = 1, ..., N$ **do**
    for $s = 1, ..., q$ **do**
      $F_{i,s}^{(t+1)} = \alpha \sum_{j \neq i} W_{i,j} F_{j,s}^{(t)} + (1 - \alpha) Y_{i,s}$
    end for
    $\delta = \max\{\delta, \max_s |P_{i,s}^{(t+1)} - P_{i,s}^{(t)}|\}$
  end for
  $t = t + 1$
end while
Saída: $\mathbf{P}^{(t)}$

2.2.2 Construção de similaridades

Nos estudos realizados por de Sousa, Rezende e Batista (SOUSA; REZENDE; BATISTA, 2013) foi evidenciado que métodos semi-supervisionados dependem significativamente da construção de G_D e $\mathbf{W}$, assim como da escolha de parâmetros utilizada tanto nesta etapa como na de inferência, o que é coerente com a formulação do problema de transdução baseada em grafos, que consiste da realização de inferência sobre as estruturas que descrevem as similaridades entre instâncias de D.

Cada vértice de G_D corresponde a uma instância de D, enquanto as arestas descrevem as relações entre pares das mesmas. A primeira etapa da construção de uma matriz de similaridade passa então pela construção de G_D, equivalente a determinar as entradas não-nulas de $\mathbf{W}$, o que pode ser feito utilizando um critério de corte (isto é, se a distância entre os atributos de duas instância é menor que um limite ϵ, então a similaridade é não-nula) ou escolhendo os k vizinhos mais próximos de uma instância, que é a abordagem mais utilizada (ENGELEN; HOOS, 2020).

Finalizada a construção de G_D, resta então determinar os pesos atribuidos a cada entrada não-nula de $\mathbf{W}$. Porém, o processo pode se dar também de forma inversa: primeiro se constrói uma matriz $\mathbf{K}$ na qual $K_{i,j} = K(x_i, x_j)$ que então passa por um processo de esparsificação(SOUSA; REZENDE; BATISTA, 2013), de forma que a matriz de similaridade $\mathbf{W}$ é obtida através da remoção das entradas menos relevantes de $\mathbf{K}$ (ENGELEN; HOOS, 2020).

Destacam-se então três métodos para construção da matriz de similaridade

- ***Clustering with Adaptive Neighbors* (CAN)**: consiste de um método que transforma o cálculo de similaridades em um problema de otimização (NIE; WANG; HUANG, 2014) cuja função a ser minimizada é, para cada $(x_i, y_i) \in D$

$$f_i(\mathbf{W}_i) = \sum_j (||x_i - x_j||^2 W_{i,j} + \gamma W_{i,j})^2 \tag{2.5}$$

  sujeita às restrições

$$W_{i,j} \in [0, 1] \text{ e } \sum_j W_{i,j} = 1 \tag{2.6}$$

  e γ é uma constante de normalização. Ao impor-se a condição que cada x_i possua apenas k vizinhos tais que $W_{i,j} > 0$, chega-se à solução (MA; XIAO; DENG, 2020)

$$W_{i,j} = \frac{d^*_{i,k+1} - d^*_{i,j}}{k d^*_{i,k+1} - \sum_l^k d^*_{l,j}} \tag{2.7}$$

  na qual $d^*_{i,j} = ||x_i - x_j||$ porém ordenada de forma crescente para que $d^*_{i,j} \leq d^*_{i,j+1}$ para todo $j = 1, ..., N$.

- ***Probabilistic Nearest Neighbors* (PNN)**: é uma modificação do CAN na qual as similaridades são normalizadas pela regra *min-max* e é imposta a condição que para cada ponto de D existam apenas k vizinhos tais que $W_{i,j} \geq 0$ (MA; XIAO; DENG, 2020). A solução resulta em similaridades da forma

$$W_{i,j} = \frac{d^*_{i,k+1} - d^*_{i,j}}{d^*_{i,k+1} - d^*_{i,1}}. \tag{2.8}$$

- ***Radial basis function* (RBF)**: sendo uma das abordagens mais utilizadas para construção de similaridades(ENGELEN; HOOS, 2020), esta consiste em utilizar como *kernel* uma função gaussiana

$$K_{i,j} = \exp\left\{\frac{-||x_i - x_j||^2}{2\sigma^2}\right\}, \tag{2.9}$$

cujo único parâmetro é a variância σ, que pode ser ajustada por (SOUSA; REZENDE; BATISTA, 2013)

$$\sigma = \frac{1}{3N}\sum_i d^*_{i,k+1}, \tag{2.10}$$

onde N é o número de instâncias no conjunto de dados.

Introduzido o aprendizado semi-supervisionado e as particularidades de métodos transdutivos, a discussão segue adiante introduzindo o modelo Potts, que corresponde a uma abordagem probabilística ao problema de transdução.

3 Modelo Potts

Assim como outras abordagens de aprendizado de máquina, a classificação transdutiva (semi-supervisionada) pode ser associada a uma função custo (ENGELEN; HOOS, 2020). Suponha uma função $H(\mathbf{s})$, onde $\mathbf{s} = (s_1, ..., s_N)$ é um vetor aleatório tal que $s_i \in \{1, ..., q\}$ para $i = 1, ..., N$, sendo q o número de classes possíveis no problema e N o número de instâncias no conjunto de dados D, de forma que cada variável s_i representa a classe atribuída a um elemento (x_i, y_i) de D.

A função $H(\mathbf{s})$ é então caracterizada por um parte *interativa* que descreve a tendência de variáveis associadas a pontos mais similares a pertecerem a uma mesma classe e um segundo termo de *campo* (ou *evidência*), que descreve o conhecimento acerca dos pontos em D_l (ENGELEN; HOOS, 2020). Uma função custo que satisfaz esses requisitos é da forma

$$H(\mathbf{s}) = -\sum_{i<j} J_{i,j}\delta(s_i, s_j) - \sum_{i,c} h_{i,c}\delta(s_i, c), \tag{3.1}$$

onde

$$\delta(a, b) = \begin{cases} 1, & \text{se } a = b, \\ 0, & \text{se } a \neq b. \end{cases} \tag{3.2}$$

A matriz $\mathbf{J}_{N\times N}$ está associada à similaridade $\mathbf{W}$ e descreve a propriedade de agrupamento entre as variáveis. Já a matriz $\mathbf{h}_{N\times q}$ descreve a tendência de variáveis previamente classificadas a manterem-se em sua classificação original.

O modelo Potts (WU, 1982) corresponde então à distribuição de probabilidade

$$p(\mathbf{s}) = \frac{1}{Z}\exp\{-H(\mathbf{s})\}, \text{ onde } Z = \sum_{\mathbf{s}} \exp\{-H(\mathbf{s})\}. \tag{3.3}$$

Z é também conhecida como função de partição em textos de Mecânica Estatística.

A distribuição na Equação 3.3 possui a propriedade desejada de atribuir maiores probabilidades a configurações de $\mathbf{s}$ com menor custo. No entanto, o resultado da inferência dependerá fortemente da escolha dos parâmetros de interação e evidência para o modelo.

3.1 Inferência com o modelo Potts

Começando pela matriz de interação $\mathbf{J}$, nota-se que para interações positivas ($J_{i,j} > 0$) o modelo deverá favorecer configurações nas quais as variáveis de um par (s_i, s_j) estão no mesmo estado. Em contrapartida, interações negativas atribuem maiores probabilidades a estados nos quais $s_i \neq s_j$.

Considerando a hipótese de agrupamento discutida anteriormente, nota-se que, para atingir o objetivo desejado de classificação transdutiva as interações não podem ser todas negativas, uma vez que isso levaria a um comportamento de repulsão generalizada entre as variáveis. No entanto, com apenas interações positivas a inferência pode levar a resultados corretos caso as interações descrevam apropridamente as similaridades entre os pontos, como foi explorado no algoritmo de agrupamento superparamagnético (BLATT; WISEMAN; DOMANY, 1996).

No trabalho de Blatt, Wiseman e Domany, a função custo possui apenas a parte interativa, visto que é um método de agrupamento no qual $D_l = \emptyset$. Nesse trabalho (e nos posteriores) vê-se que a matriz de interação não descreve puramente a similaridade entre as variáveis, mas também a intensidade de interação entre as mesmas. Em particular, as interações são da forma $J_{i,j} = \beta\omega_{i,j}$, na qual $\omega_{i,j}$ está associada à similaridade do par de variáveis e $\beta \geq 0$ descreve a intensidade de interação, sendo o parâmetro a ser ajustado na etapa de inferência.

No caso em que as interações são positivas, o modelo é dito *ferromagnético* e pode-se adotar $\boldsymbol{\omega} = \mathbf{W}$.

Embora a utilização de interações estritamente negativas (*antiferromagnéticas*) não levem aos resultados desejados de agrupamento e classificação, a coexistência de interações positivas e negativas apropriadamente construídas permite esse tipo de aprendizado (SHI; LIU; ZHANG, 2018). Nesse caso, o modelo será dito um *vidro de spin*, e a matriz $\boldsymbol{\omega}$ pode ser construída como

$$\omega_{i,j} = W_{i,j} - \bar{\omega}, \text{ onde } \bar{\omega} = \frac{1}{N^2}\sum_{i,j} W_{i,j}. \tag{3.4}$$

Tomando agora a matriz de evidência $\mathbf{h}$, a análise da mesma é similar à da matriz de interação: para $h_{i,k} > 0$ o modelo favorece estados nos quais $s_i = k$ e, caso $h_{i,k} < 0$, configuraçãoes com $s_i \neq k$ serão favorecidas. Portanto, esses campos devem agir sobre os elementos de D_l de forma a introduzir uma tendência nas variáveis associadas a tais elementos que os faz permanecer no rótulo já conhecido (WANG et al., 2007; LI; WANG, 2016). No caso do modelo Potts, uma configuração possível da evidência é

$$h_{i,k} = \begin{cases} 1, \text{ se } y_i = k \\ 0, \text{ se } y_i \neq k \text{ ou } (x_i, y_i) \notin D_l. \end{cases} \tag{3.5}$$

Conectando essa definição com os estudos sobre agrupamento com o modelo Potts, vê-se que este é um bom exemplo do que foi apresentado no capítulo anterior: a classificação semi-supervisionada transdutiva corresponde a um algoritmo de agrupamento com informação adicional acerca das classes de alguns pontos. No caso de agrupamento, a matriz $\mathbf{h}$ é nula e os grupos são determinados pelas correlações entre pares de variáveis

$$C_{i,j} = \langle \delta(s_i, s_j) \rangle_p \tag{3.6}$$

sendo que a notação

$$\langle A \rangle_b = \sum_x A(x) b(x) \tag{3.7}$$

denota a média de uma função $A(x)$ em relação a uma distribuição de probabilidade $b(x)$. Já em uma situação de classificação, os rótulos podem ser determinados pelo estado mais provável de cada variável (WANG et al., 2007; LI; WANG, 2016):

$$y_i = arg\max_s \langle \delta(s_i, s) \rangle_p. \tag{3.8}$$

3.2 Aproximações de campo médio

Da expressão Equação 3.3 nota-se que as médias de interesse no procedimento de inferência podem ser calculadas a partir de derivadas parciais da energia livre $F_H = -\ln Z$ em relação aos parâmetros do modelo:

$$C_{i,j} = -\frac{\partial F_H}{\partial J_{i,j}} \tag{3.9}$$

e

$$p_{i,s} = \langle \delta(s_i, s) \rangle_p = -\frac{\partial F_H}{\partial h_{i,s}}. \tag{3.10}$$

Essa propriedade é particularmente útil ao notar que a outra alternativa para calcular essas propriedades envolveria geralmente somar sobre as q^N configurações do modelo. Portanto, médias que podem ser calculadas das derivadas parciais de F_H são computacionalmente viáveis uma vez que seja conhecida uma expressão analítica ou uma aproximação da energia livre.

Uma forma de contornar estas limitações foi empregada no trabalho inaugural sobre agrupamento superparamagnético (BLATT; WISEMAN; DOMANY, 1996) e envolve a utilização de métodos de Monte Carlo, no qual o objetivo é produzir amostras pseudoaleatórias da distribuição $p(\mathbf{s})$ e utilizá-las no cálculo das médias de interesse. Posteriormente, uma forma determinística de atacar o problema foi estudada utilizando o algoritmo de *Belief Propagation* (OTT; STOOP, 2006). Tal algoritmo foi proposto na década de 1980 (PEARL, 1982) e posteriormente identificado como uma aproximação para F_H quando $\mathbf{J}$ é esparsa (YEDIDIA; FREEMAN; WEISS, 2000). Em particular, essa aproximação é exata quando o grafo de similaridades é uma árvore. Recentemente, esse algoritmo foi aplicado nos problemas de agrupamento semi e não-supervisionado utilizando similaridades de vidro de spin (SHI; LIU; ZHANG, 2018).

No contexto de classificação semi-supervisionada, destaca-se um trabalho com o modelo Potts no qual também foi utilizado um método de Monte Carlo (GETZ; SHENTAL; DOMANY, 2005). Posteriormente houveram novos desenvolvimentos utilizando técnicas de aproximação determinísticas porém fazendo uso do modelo Ising (WANG et al., 2007; LI; WANG, 2016), que corresponde ao modelo Potts no caso em que $q = 2$.

As aproximações determinísticas mencionadas anteriormente são conhecidas como *aproximações de campo médio* e por vezes podem ser resumidas à busca por uma aproximação da energia livre (YEDIDIA, 2001).

Enquanto métodos de Monte Carlo possuem como principal desvantagem o uso de números pseudoaleatórios, métodos de campo médio podem ser custosos computacionalmente: *belief propagation* (YEDIDIA, 2001) e a aproximação TAP (Thouless-Anderson-Palmer) adaptativa (LI; WANG, 2016), por exemplo, possuem complexidade $O(N^3)$. Logo, é de extremo interesse o estudo de aproximações cuja complexidade não limite a aplicabilidade em conjuntos de dados reais (ENGELEN; HOOS, 2020; LI; WANG, 2016).

3.2.1 A aproximação variacional

Será descrita agora uma aproximação de campo médio para o modelo Potts análoga à utilizada em (WANG et al., 2007) para o modelo Ising. Nessa abordagem, o objetivo é minimizar a divergência de Kullback-Leibler (KL)

$$D_{KL}(\phi||p) = \sum_{\mathbf{s}} \phi(\mathbf{s};\theta) \ln \frac{\phi(\mathbf{s};\theta)}{p(\mathbf{s})}, \tag{3.11}$$

onde $\phi(\mathbf{s},\theta)$ é uma distribuição parametrizada por θ e a minimização de D_{KL} em relação a tal parâmetro fornece uma aproximação de $p(\mathbf{s})$. Inserindo a Equação 3.3 na Equação 3.11 e notando que $D_{KL}(\phi||p) \geq 0$ (OPPER; SAAD, 2001) chega-se à desigualdade

$$\langle H \rangle_\phi - S[\phi] \geq F_H, \tag{3.12}$$

onde $S[\phi] = -\sum_{\mathbf{s}} \phi(\mathbf{s},\theta) \ln \phi(\mathbf{s},\theta)$ é a entropia de Shannon. Da Equação 3.12 é possível identificar o propósito de minimizar $D_{KL}(\phi||p)$: o termo à esquerda, que carrega a dependência de ϕ, consiste de um limite superior para a energia livre original F_H e pode ser interpretado como uma aproximação da mesma.

Tomando uma matriz $\boldsymbol{\theta}_{N\times q}$, e $\phi(\mathbf{s},\boldsymbol{\theta})$ uma distribuição na qual as componentes de $\mathbf{s}$ são independentes, visa-se então minimizar a divergência KL para uma distribuição da forma

$$\phi(\mathbf{s},\boldsymbol{\theta}) = \prod_i \phi_i(s_i,\boldsymbol{\theta}_i) = \prod_i \frac{1}{Z_i} \exp\{\theta_{i,s_i}\}, \text{ onde } Z_i = \sum_k \exp\{\theta_{i,k}\}. \tag{3.13}$$

Através da minimização de Equação 3.11 é obtida a expressão ótima dos parâmetros:

$$\theta_{i,s_i} = \sum_{j\neq i} J_{i,j}\phi_j(s_i) + h_{i,s_i} \text{ para } i = 1,...,N \text{ e } s_i = 1,...,q \tag{3.14}$$

e, no Algoritmo 3.2.1 está exposto o procedimento iterativo para solução numérica das equações acima. Uma vez que uma solução é obtida, as variáveis de D são classificadas como

$$y_i = arg\max_{s_i} \phi(s_i). \tag{3.15}$$

A aproximação variacional como descrita acima é também conhecida como *Naive Mean Field Theory*(NMF) e pode ser generalizada de diferentes maneiras, como a expansão de Plefka (OPPER; SAAD, 2001; GEORGES; YEDIDIA, 1991) ou através de aproximações da energia livre de Bethe (SHI; LIU; ZHANG, 2018). Diantes destas abordagens, NMF pode ser vista como a mais simples entre diversas aproximações para a distribuição de probabilidade do modelo Potts.

Algoritmo 3.2.1 Solução iterativa das equações de campo médio

Entrada: $\boldsymbol{\phi}^{(0)}$, $\mathbf{J}$, $\mathbf{h}$, ϵ, t_{max}
$t \leftarrow 0$, $\delta \leftarrow \epsilon$
while $t < t_{max}$ e $\delta \geq \epsilon$ **do**
  $\delta \leftarrow 0$
  for $i = 1, ..., N$ **do**
    for $k = 1, ..., q$ **do**
      $$\phi_i^{(t+1)}(k) \leftarrow \frac{\exp\left\{\theta_{i,k}^{(t)}\right\}}{\sum_l \exp\left\{\theta_{i,l}^{(t)}\right\}}$$
    end for
    $\delta \leftarrow \max\{\delta, \max_k |\phi_i^{(t+1)}(k) - \phi_i^{(t)}(k)|\}$
  end for
  $t \leftarrow t + 1$
end while
Saída: $\boldsymbol{\phi}^{(t)}$

3.3 Ajuste do parâmetro β

Como discutido previamente, o principal problema a ser resolvido é ajustar o parâmetro β. Em problemas de agrupamento (caso $\mathbf{h} = \mathbf{0}$) existem diversas abordagens tanto para modelos ferromagnéticos (BLATT; WISEMAN; DOMANY, 1996; OTT; STOOP, 2006) como em vidros de spin (SHI; LIU; ZHANG, 2018). Para o caso da transdução semi-supervisionada, no entanto, as abordagens conhecidas tratam o problema como uma extensão do problema de agrupamento (GETZ; SHENTAL; DOMANY, 2005; SHI; LIU; ZHANG, 2018).

No entanto, já foi reportado que o valor optimal β^* decresce em função da quantidade pontos previamente classificados (GETZ; SHENTAL; DOMANY, 2005). Um procedimento de ajuste específico para o problema de transdução deve então ser capaz de capturar tal comportamento.

Em particular, no caso ferromagnético, o valor ótimo de β é apontado como pertencendo a uma faixa de valores denominada a *fase superparamagnética* do modelo (BLATT; WISEMAN; DOMANY, 1996; GETZ; SHENTAL; DOMANY, 2005). Nessa fase, diferentes valores de β estão associados a diferentes estruturas de *cluster* presentes no conjunto, de forma que a escolha de um valor de β implica escolher uma dentre essas

possíveis estruturas. Nesse caso, é sugerido que uma estimativa de β^* é dada por (BLATT; WISEMAN; DOMANY, 1996)

$$\beta_{spm} = 4 \exp\{1/2\} \ln\{1 + \sqrt{q}\}. \tag{3.16}$$

O procedimento acima é oriundo de estudos do modelo Potts aplicado em tarefas de agrupamento. Como discutido previamente, é esperado que a classificação semi-supervisionada seja possível somente em situações nas quais a informação estrutural do conjunto esteja associada aos rótulos a serem atribuídos a cada instância.

Visto que β regula a intensidade de interação entre pares de variáveis, tal parâmetro está associado à influência das propriedades de agrupamento do modelo. Assim, $\beta = 0$ corresponde ao caso em que a informação estrutural é totalmente ignorada, análogo à percepção de que existem N *clusters* no conjunto. Conforme $\beta \to \infty$ a quantidade de agrupamentos percebida diminui, chegando ao caso extremo em que todas a instâncias possuirão o mesmo rótulo (ou seja, pertecerão a um mesmo grupo).

Em diversos trabalhos de SSL é comum a matriz de interação ser normalizada de forma que, para $i = 1, ...N$ (ZHU; GHAHRAMANI, 2002; SOUSA; REZENDE; BATISTA, 2013)

$$\sum_{j \neq i} \omega_{i,j} = 1. \tag{3.17}$$

Nesse caso, obtemos que

$$\beta \sum_{j \neq i} \omega_{i,j} \phi_j(s_i) < \theta_{i,s_i} \tag{3.18}$$

sempre que $\beta \leq 1$. Conectando essa discussão com a Equação 3.14, conclui-se que as probabilidades obtidas nessa situação são mais influenciadas pela informação *a priori* do que pela informação angariada através do conhecimento estrutural do conjunto. Nota-se também que $\beta_{spm} > 1$.

No entanto, como já discutido, $\beta > 1$ não pode ser uma condição necessária (muito menos suficiente) para o sucesso do algoritmo visto que o crescimento inderteminado desse parâmetro leva à inferência de um único *cluster* no conjunto. Porém, como β^* decresce com a aquisição de mais informação acerca dos rótulos iniciais, espera-se que $\beta^* \to 1$ à medida que aumenta-se o número de elementos em D_l.

4 Metodologia

O principal objetivo deste trabalho é estudar as possibilidade de ajuste do parâmetro β do modelo Potts na tarefa de classificação transdutiva. Pelo discutido anteriormente, dois valores de β^* chamam a atenção:

$$\beta^* = \beta_{spm} \quad \text{e} \quad \beta^* = 1. \tag{4.1}$$

Visto que a matriz de interação $\boldsymbol{\omega}$ é obtida da matriz de similaridade (que por sua vez é construída a partir dos métodos apresentados na subseção 2.2.2) e a matriz de campos $\mathbf{h}$ é definida como na Equação 3.5, β é o único parâmetro a ser ajustado. Assim, experimentos são realizados sobre três conjuntos de dados bidimensionais (Figura 2) e em outros seis conjuntos de *benchmark* cujos *ground truths* são definidos de forma a comparar as abordagens de ajuste expostas acima.

A construção de similaridades é feita da seguinte maneira: fixado um número de vizinhos mais próximos k uma matriz de similaridade $\hat{\mathbf{W}}$ é obtida utilizando PNN ou RBF para pesagem das arestas. Nesse processo são considerados que cada nó possui um igual número de vizinhos, o que leva a similaridades assimétricas. De forma a obter uma similaridade simétrica $\mathbf{W}$ toma-se (SOUSA; REZENDE; BATISTA, 2013)

$$W_{i,j} = \min\{\hat{W}_{i,j}, \hat{W}_{j,i}\} \tag{4.2}$$

e em seguida liga-se cada possível nó isolado a seu vizinho mais próximo, atribuindo a esta aresta o maior peso calculado previamente em $\hat{\mathbf{W}}$ (SOUSA; REZENDE; BATISTA, 2013). Os métodos são então avaliados para $k = 2, 3, ..., 20$. Como métrica, utiliza-se sempre a distância euclideana. Ao final desse processo, para um dado k os algoritmos de propagação recebem como entrada a matriz de similaridade normalizada

$$\boldsymbol{\omega} = \mathbf{S}^{-1/2}\mathbf{W}\mathbf{S}^{-1/2}, \tag{4.3}$$

onde $\mathbf{S}$ é a matriz diagonal tal que

$$S_{i,i} = \sum_{j \neq i} W_{i,j}. \tag{4.4}$$

Nos conjunto bidimensionais, em particular, avalia-se a dependência do algoritmo 3.2.1 em relação ao parâmetro β. Fixados $r_l = 0,10$ e duas topologias de G_D nas quais $k = \log_2 N$ e $k = N-2$, respectivamente, são mensuradas a acurácia e o tempo de execução para cada valor de β no *grid* $10^{[-5:5:0,25]}$.

É necessário também avaliar o algoritmo para diversas construções de D_l, chamadas de *splits*. Para cada conjunto são gerados aleatoriamente 20 *splits* distintos sob a restrição

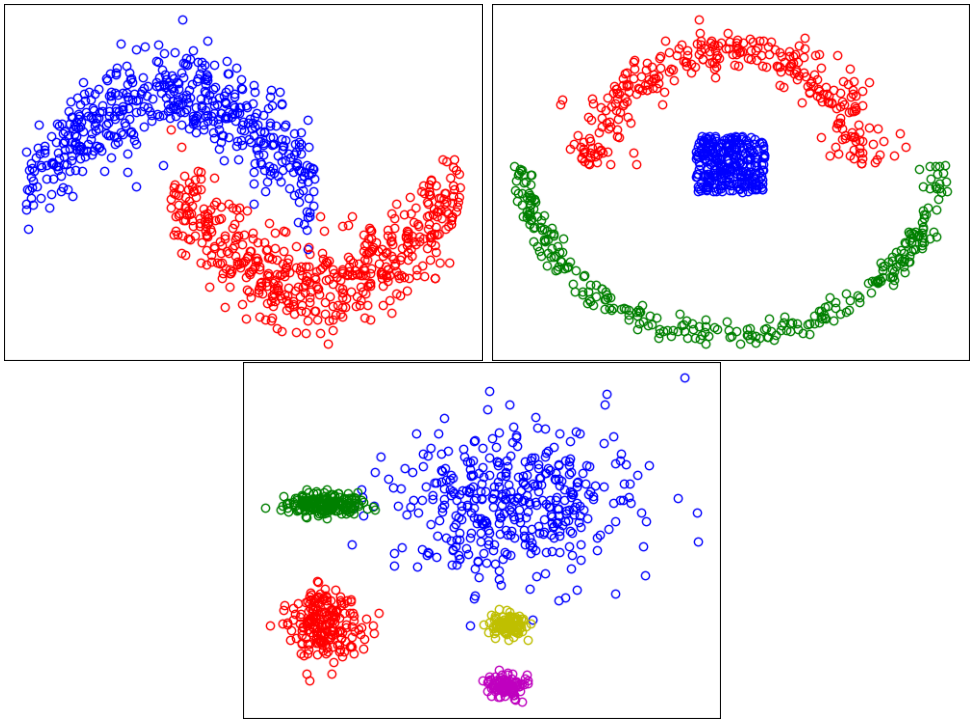

Figura 2 – Os três conjuntos artificais utilizados para avaliação do modelo Potts aplicado à classificação semi-supervisionada. Cores distintas indicam rótulos distintos. Em cima: conjunto de duas luas(esq.) e conjunto de três *clusters* (dir.); em baixo: conjunto de cinco gaussianas independentes.

de que cada um destes contenha ao menos um representante de cada classe. No caso dos conjuntos bidimensionais, estes são avaliados somente para $r_l = 0,1$, enquanto que, para os demais, toma-se o *grid* $r_l \in [0,02 : 0,2 : 0,02]$.

Os conjuntos bidimensionais são construídos de forma a apresentar uma alta separabilidade entre as classes, o que possibilita avaliar os algoritmos em uma situação próxima ideal, na qual há pouca ou nenhuma ambiguidade entre um *cluster* e uma *classe*. É também nesse tipo de configuração em que a acurácia é uma medida significativa do desempenho de diferentes construções e métodos. Os ditos conjunto são:

- **Duas luas:** consiste de 1000 instâncias divididas em duas regiões não-convexas balanceadas e com formas similares.
- **Três *clusters*:** contém 900 instâncias divididas em três regiões, sendo duas convexas e uma não-convexa, balanceadas e com formas distintas.
- **Cinco gaussianas:** apresenta 1000 pontos extraídos de cinco gaussianas desbalanceadas com médias e variâncias distintas.

Quanto aos conjuntos de *benchmark* disponíveis na literatura, estes são

- **Digit1** é uma coleção de 1500 imagens artificiais do dígito 1 divididas em duas classes. Tais imagens são preprocessadas em instâncias de 241 dimensões de forma a obscurecer o conteúdo da imagem. Tais dados estão próximos a uma variedade de baixa dimensão mas não exibem uma estrutura de *cluster* bem definida (CHAPELLE; SCHOLKOPF; ZIEN, 2006).

- **COIL** é um subconjunto de COIL-100 contendo seis classes e 1500 instâncias correspondentes a diversos objetos, originalmente divididos em 24 classes. Tais imagens foram preprocessadas de forma a reduzir seu tamanho de 128×128 para 16×16 pixels e seu conteúdo foi obscurecido da mesma forma que para o conjunto Digit1 (CHAPELLE; SCHOLKOPF; ZIEN, 2006). O conjunto dos atributos também possui dimensão 241.

- **Mnist** é um subconjunto de MNIST (LECUN et al., 1998) que consiste de 5000 imagens do conjunto original divididas em 10 classes (RASCHKA, 2018). Tais imagens correspondem a escritas cursivas dos dígitos de 0 a 9 em um fundo branco com dimensões de 28×28 pixels.

- **USPS** (HULL, 1994) também é um conjunto de imagens de escritas cursivas dos digitos de 0 a 9, porém contendo 9298 imagens com dimensões de 16×16 pixels divididas em 10 classes.

- **Phoneme** é um conjunto do respositório KEEL (ALCALÁ-FDEZ et al., 2011) no qual cada uma das 5404 instâncias descreve um de dois tipos de sons (oral ou nasal) através de cinco atributos. As duas classes estão desbalanceadas em proporção de aproximadamente 2/5 entre sons orais e nasais.

- **Texture** é um conjunto do repositório KEEL (ALCALÁ-FDEZ et al., 2011) no qual a cada instância é atribuída uma dentre onze texturas possíveis. Cada instância corresponde a um padrão descrito por quarenta atributos extraídos a partir da estimação de momentos de quarta ordem. Este conjunto conta com 5500 padrões distribuídos entre as textura disponíveis.

Em todos os conjuntos, NMF é comparado a LP e LGC. Quanto aos algoritmos de propagação são utilizados em todos os métodos $t_{max} = 10^4$, $\epsilon = 10^{-3}$ e como condição inicial é adotada a distribuição uniforme

$$F_{i,s}^{(0)} = P_{i,s}^{(0)} = \phi_i^{(0)}(s) = \frac{1}{q} \tag{4.5}$$

para $i = 1, ..., N$ e $s = 1, ..., q$. Também tomamos $\alpha = 0,99$ para o algoritmo LGC, como foi feito no artigo original (ZHOU et al., 2004).

Como métricas de avaliação dos algoritmos são utilizadas a acurácia, definida como a taxa de pontos corretamente classificados e o tempo de execução, de forma a avaliar o desempenho computacional dos algoritmos.

4.1 Implementação

Os algoritmos de propagação 2.2.1, 2.2.2 e 3.2.1 são implementados na linguagem *C* (RITCHIE; KERNIGHAN; LESK, 1988) e então passados por uma interface em Cython (BEHNEL et al., 2010) que permite chamar tais funções em Python (ROSSUM; DRAKE, 2009).

Esse procedimento permite, via Python, chamar códigos que fazem uso de paralelização *multithread*, nesse caso utilizando OpenMP (DAGUM; MENON, 1998) e obter um significativo ganho de performance, a depender da quantidade de *threads* disponíveis. Tal melhora é isomorfa aos três métodos pois a paralelização é feita sobre o primeiro *loop for* de cada algoritmo. Dessa forma, a cada iteração do *loop* uma *thread* se encarrega de atualizar a função atribuída a um nó de G_D.

Os experimentos foram realizados em um processador Intel i5-1135G7 com 8GB de RAM utilizando todas as 8 *threads* disponíveis. Os códigos utilizados estão disponíveis em https://github.com/boureau93/ssl-nmf.

5 Resultados e discussão

De forma a estabelecer uma compreensão mais apurada dos métodos em questão, estes são primeiro avaliados nos conjuntos bidimensionais e em seguida passa-se a uma análise nos conjuntos de alta dimensão.

5.1 Conjuntos bidimensionais

Na Figura 3 estão dispostos os resultados utilizando o algoritmo 3.2.1 em função de β para diferentes topologias do grafo de similaridade utilizando as similaridades PNN e RBF. Primeiramente, é notável a diferença de comportamento do algoritmo em topologias mais esparsas ($k = \log_2 N$) e mais densamente conectadas ($k = N - 2$). No primeiro caso, a região de ótima acurácia está associada a menores valores da energia interna, enquanto o segundo caso só pode ser considerado solidamente bem sucedido na tarefa de aprendizado na região de menor energia, na qual $\beta << 1$.

Esse comportamento decorre do fato de que, à medida que a conectividade do grafo de similaridade aumenta, a topologia do grafo torna-se menos relevante para a descrição das classes, de forma que os pesos tornam-se o fator mais relevante à descrição estrutural de um conjunto de dados. Assim, para similaridades densamente conexas, conforme $\beta \to \infty$, a distinção feita pelos pesos desaparece devido à alta correlação entre as variáveis do modelo, levando a uma fase ferromagnética na qual o modelo detecta um único *cluster*.

No caso de similaridades esparsas, NMF apresenta a existência de três *plateaus* de acurácia, sendo os dois primeiros indistinguíveis energeticamente, enquanto o terceiro apresenta uma forte associação com uma região de menor energia interna. Esse comportamento é próprio do caso em que a topologia de G_D é capaz de separar suficientemente bem as classes, impedindo a manifestação de uma fase ferromagnética global, como ocorre no caso denso.

Quanto ao tempo de execução, similaridades densamente conectadas mostraram um perfil mais bem comportado no tempo de execução, à excessão do conjunto de três *clusters*, no qual o comportamento do tempo de execução não parece drasticamente afetado pela topologia da representação. No entanto, nota-se como o algoritmo torna-se mais lento na região de menor energia, que é um fenômeno conhecido como desaceleramento crítico (*critical slowing down*) (OTT; STOOP, 2006), apresentado por modelos como o de Potts nas vizinhanças de uma transição de fase.

Para β suficientemente grande é observado que o algoritmo tem sua convergência impossibilitada, o que associa-se à precisão numérica do computador utilizado nos

experimentos. Isso deve-se ao fato da outra alternativa de explicar tal comportamento ser pela divergência das próprias equações de campo médio, o que não ocorre nesse caso devido à convexidade da divergência de Kullback-Leibler (BERTSEKAS; HAGER; MANGASARIAN, 1998).

Quanto ao ajuste de parâmetros, nota-se que $\beta = 1$ está associado a uma região levemente suboptimal do modelo em termos de acurácia, assim como uma alta energia. β_{spm}, pelo contrário, está associada à região de pior desempenho computacional do algoritmo, que ocorre em baixas energias. Portanto, o parâmetro β define um *trade-off* entre acurácia e tempo de execução no caso esparso, enquanto no caso de uma representação densa a região de melhor desempenho é também aquela na qual a acurácia é mais estável em torno de seu valor optimal.

A Figura 4 mostra o comportamento dos quatro métodos estudados em função do número de vizinhos mais próximos k em uma região na qual G_D ainda pode ser considerado esparso. É notável que para valores mais elevados de k os três algoritmos operam de forma similar em termos de acurácia, com a abordagem $\beta = 1$ destacando-se como a mais interessante em similaridades mais conectadas devido a seu desempenho computacional superior, o que é independente da construção de G_D.

Quanto às diferenças entre PNN e RBF, destaca-se que a acurácia mostrou-se muito mais dependente da topologia do grafo do que do processo de pesagem. A principal diferença entre os dois métodos de pesagem utilizados parece manifestar-se mais fortemente nos tempos de execução, principalmente no caso $\beta = \beta_{spm}$ de NMF (Figura 4).

Em relação ao comportamento dos algoritmos em função da quantidade de pontos previamente classificados (Figura 5), nota-se que, em uma construção esparsa de G_D, o caso $\beta = 1$ é mais apropriado para configurações nas quais r_l não é proibitivamente pequeno. De fato para $r_l \geq 0,10$ a diferença de acurácia entre os algoritmos é desprezível, ao passo que a referida abordagem é seguramente mais rápida que as demais.

Ainda considerando o caso $k = \log_2 N$, para $\beta = \beta_{spm}$ pode-se garantir apenas que seu desempenho computacional é inferior ao outro ajuste de NMF (consequência do desaceleramento crítico), visto que seu tempo de execução quando comparado a LP e LGC é muito variável. Em termos de acurácia, no entanto, o mesmo é indistinguível dessas duas abordagens, tanto em função de k (Figura 4), como em função de r_l (Figura 5).

Analisa-se agora o caso de uma similaridade densamente conectada, cujos resultados estão dispostos na Figura 6. Tendo em vista a discussão realizada acerca da Figura 3, a utilização de $\beta = \beta_{spm}$ deve produzir resultados de baixa acurácia. Assim, no caso de $k = N - 2$ foram testados $\beta = 1$ e $\beta = 10^{-3}$.

Para G_D densamente conectado as abordagens utilizando NMF mostraram-se largamente mais vantajosas. A curva de acurácia apresentada pelo método LGC mostra que o

mesmo é incapaz de realizar transdução em uma topologia densa. LP apresenta certas irregularidades, mas em uma observação mais abrangente o perfil de acurácia mostra-se crescente conforme aumenta-se a quantidade de instâncias previamente classificadas(Figura 6).

Ambas abordagens de NMF possuem acurácia significativamente superior a LP e LGC em uma topologia densa, com $\beta = 10^{-3}$ mostrando-se levemente melhor quanto a essa métrica. Em termos de tempo de execução, NMF segue novamente com a abordagem de melhor tempo de execução (Figura 6), em particular $\beta = 10^{-3}$, que está situada ainda mais distante do valor crítico de β.

Ressalta-se, por fim, que a utilização de similaridades densamente conectadas não oferece vantagens em termos de acurácia quando comparada às esparsamente conectadas. De fato, como relatado em outros trabalhos , similaridades esparsas são críticas para aquisição de transdução optimal (CHONG et al., 2020; ENGELEN; HOOS, 2020). NMF, no entanto, mostra-se até aqui, sob as condições impostas, uma forma muito mais consistente de realizar transdução em similaridades densas do que LP e LGC.

5.2 Conjuntos de alta dimensão

Nos conjuntos de alta dimensão são avaliadas a acurácia e o tempo de execução dos algoritmos em função da taxa de pontos previamente classificados r_l utilizando as similaridades PNN (Figura 7) e RBF (Figura 8) com $k = \log_2 N$.

Em termos de acurácia, NMF com $\beta = \beta_{spm}$ é mais próximo da acurácia de LP e LGC na maior parte dos casos. Porém, de forma geral, esta abordagem é ou a mais custosa computacionalmente ou possui um padrão de desempenho irregular ao ponto de não ser possível afirmar se sua utilização oferece qualquer vantagem em relação aos métodos já estabelecidos.

O ajuste com $\beta = 1$, no entanto, é ao menos uma ordem de grandeza mais rápido que LP e LGC. Em situações em que a informação *a priori* é escassa ($r_l < 0,1$) a acurácia deste método é significativamente inferior à dos demais, à excessão do conjunto Phoneme.

Conforme r_l aumenta, $\beta = 1$ firma-se como uma abordagem com um melhor *trade-off* entre acurácia e tempo de execução em relação aos demais métodos. Em particular, à exceção de USPS, quando $r_l \geq 0,1$, a diferença de acurácia entre o mais rápido e o de maior acurácia é inferior a $0,05$.

O conjunto Phoneme é particularmente interessante pois é o único no qual $\beta = 1$ apresenta maior acurácia que $\beta = \beta_{spm}$. Esse comportamento é indicativo de que existe uma faixa de valores de β que comporta diversas estruturas de *cluster*, e a concordância de uma destas com o *ground truth* do conjunto não está totalmente associada a configurações de menor energia interna. Em outras palavras, os rótulos são melhores descritos por um

mínimo local da energia interna do que pelo seu mínimo global.

5.3 Sobre o caso $\beta = 1$

Frente às discussões realizadas, $\beta = 1$ mostra-se a situação mais interessante de NMF quando comparada aos métodos de *baseline* LP e LGC, principalmente devido a seu desempenho computacional, que mostrou-se consideravelmente superior ao do prévio estado da arte em todos os testes.

Os três algoritmos de propagação estudados pertencem à mesma classe de complexidade, portanto, à excessão de operações realizadas dentro dos *loops*, a principal diferença entre o tempo de execução dos mesmo deve-se a diferentes velocidades de convergência. Assim, o caso $\beta = 1$ converge mais rapidamente que os demais. Apontar as razões dessa diferença, no entanto, demanda uma avaliação mais aprofundada.

Embora tenha acurácia e tempo de execução levemente piores em relação à abordagem com $\beta = 10^{-3}$ no caso de similaridades densas, $\beta = 1$ destaca-se por operar em todas as similaridades exploradas, mostrando-se a mais robusta quanto a variações na topologia de G_D dentre todas as dinâmicas de propagação estudadas.

É importante também ressaltar que os experimentos realizados não ofereceram evidências para rejeitar a hipótese de que $\beta = 1$ é o limite inferior para um valor crítico β^* conforme $r_l \to 1$. Oferece-se mais evidências, no entanto, para a hipótese de que β^* é uma função decrescente da informação *a priori* disponível ao aprendiz, devido à alta susceptibilidade da acurácia de $\beta = 1$ em relação a r_l.

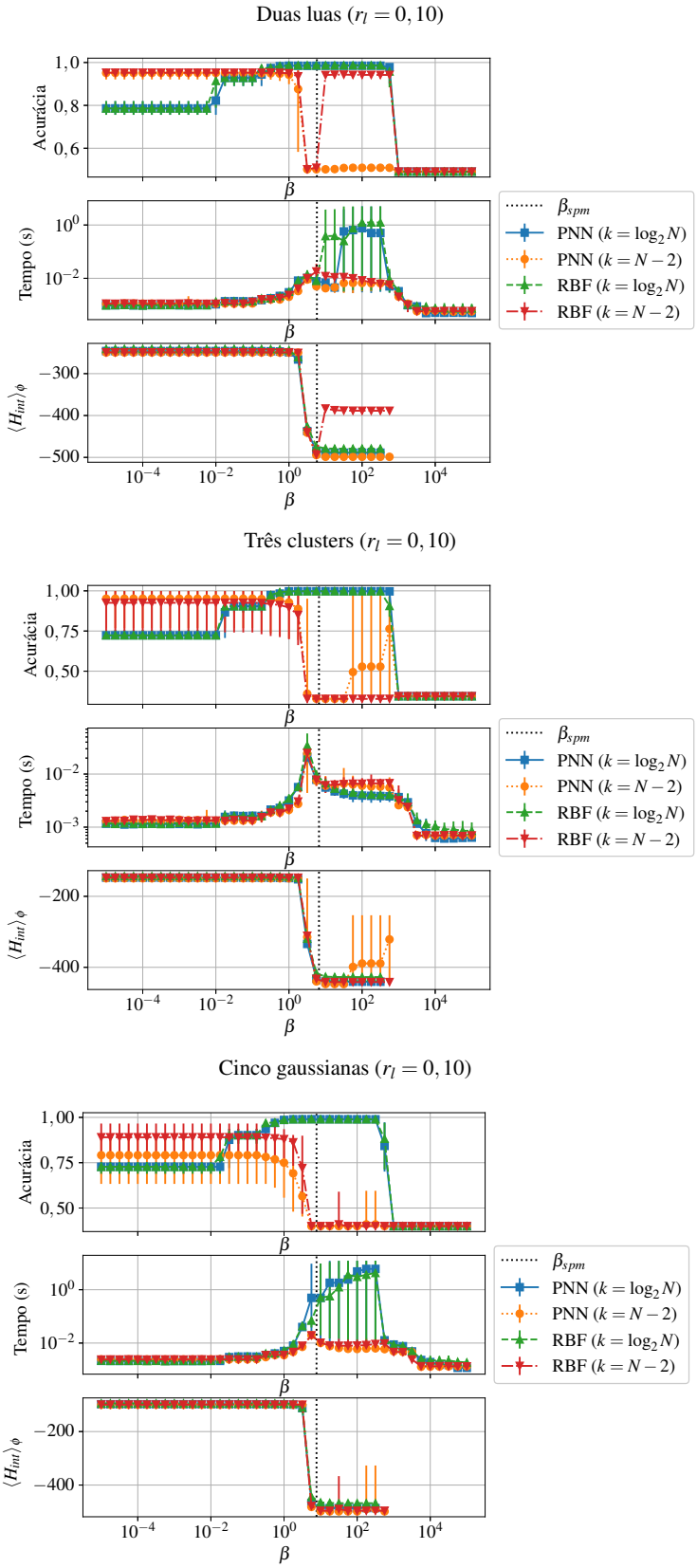

Figura 3 – Resultados da acurácia, tempo de execução e energia interna dos algoritmos NMF utilizando quatro similaridades distintas. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados.

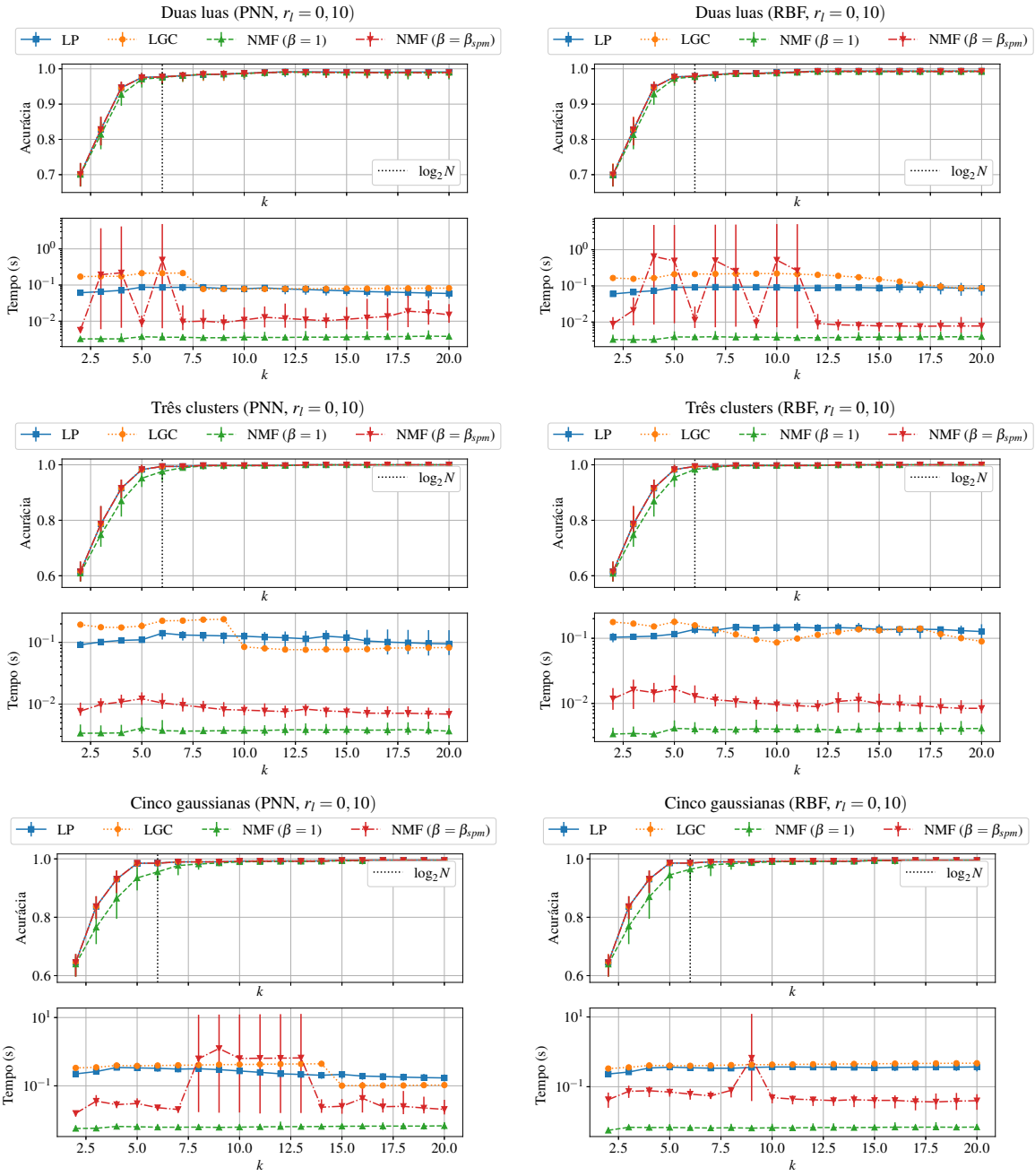

Figura 4 – Resultados da acurácia e tempo de execução em função do parâmetro k para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados.

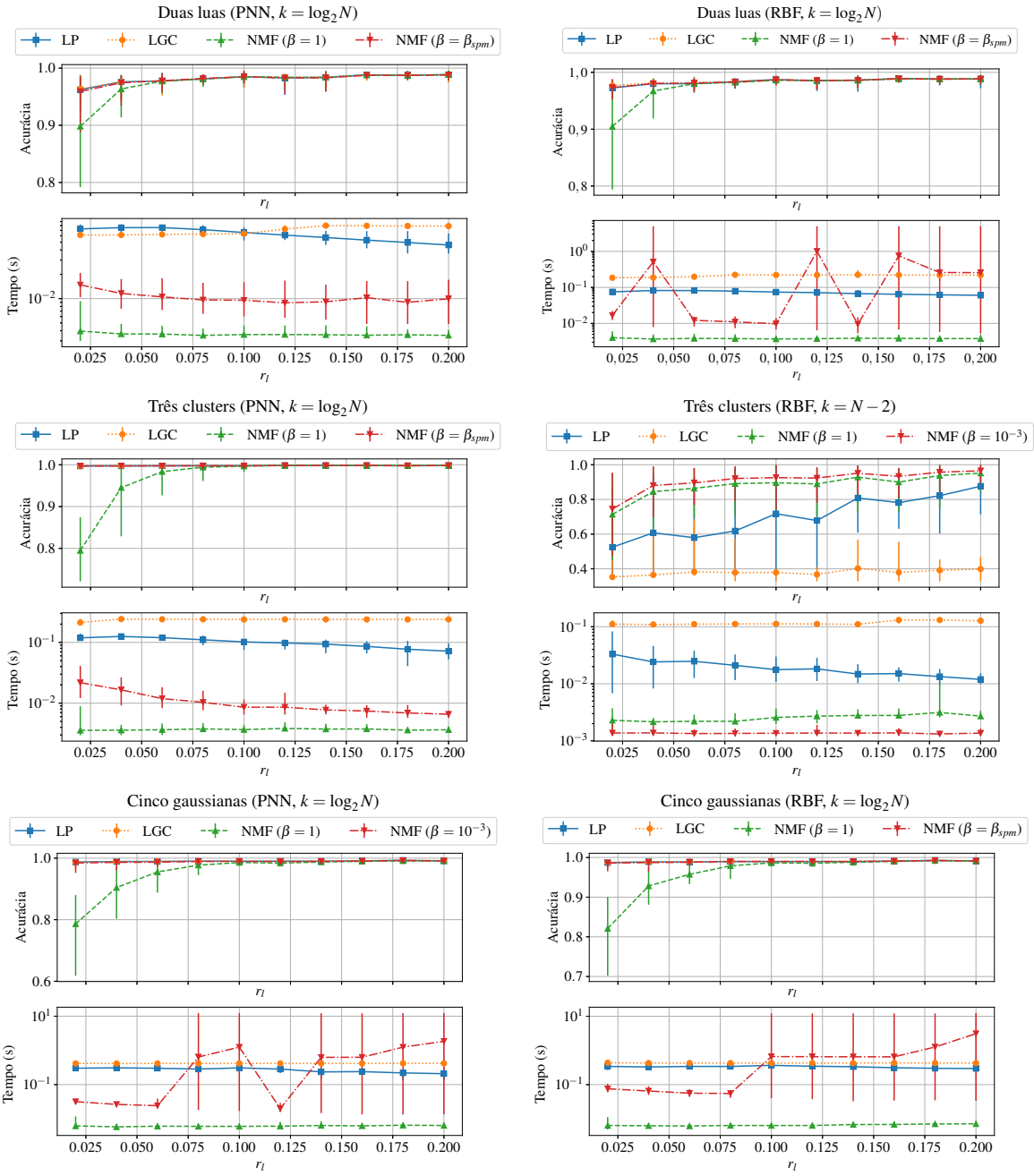

Figura 5 – Resultados da acurácia e tempo de execução em função da taxa de instâncias previamente classificadas utilizando uma similaridade esparsamente conectada para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados.

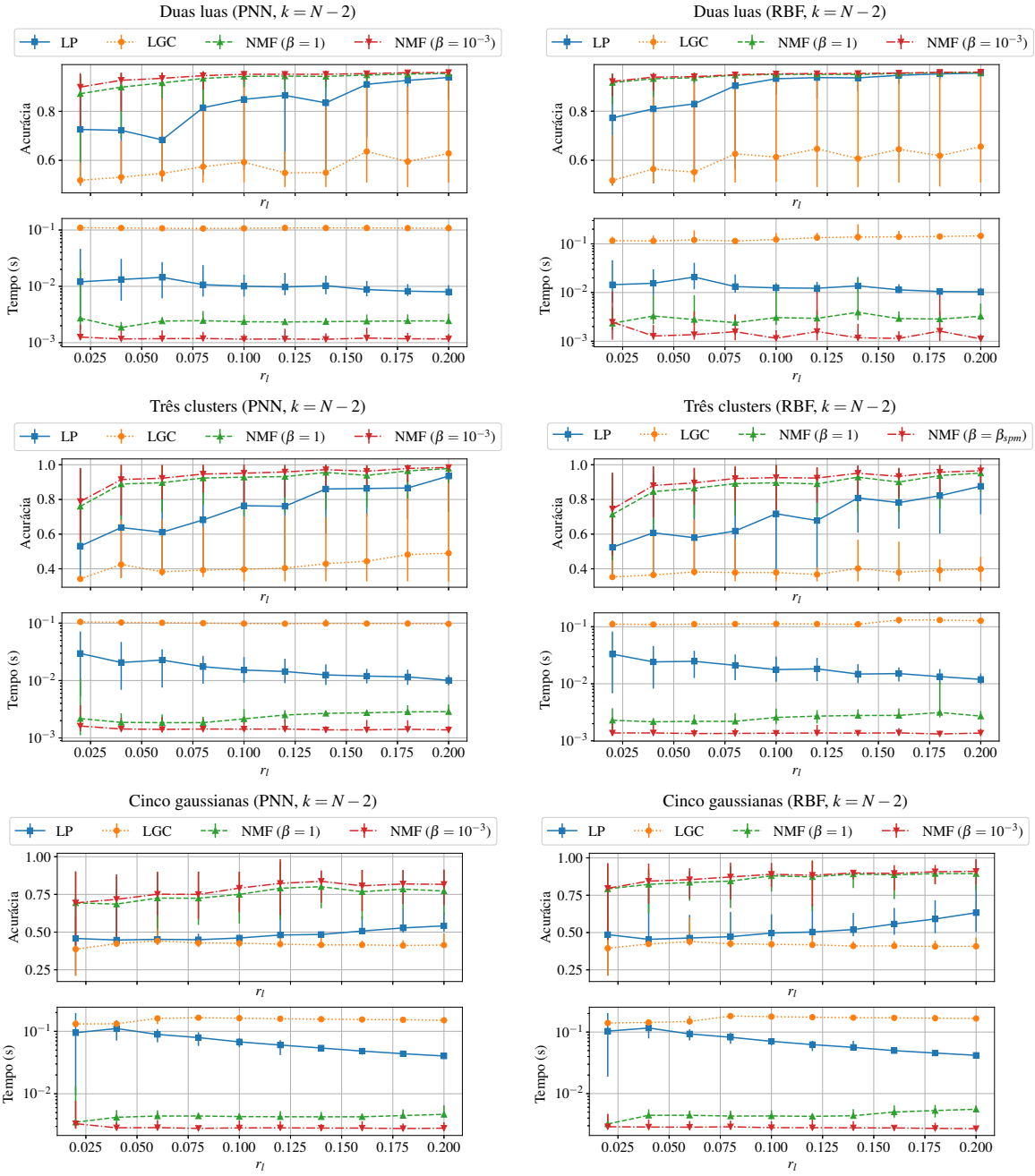

Figura 6 – Resultados da acurácia e tempo de execução em função da taxa de instâncias previamente classificadas utilizando uma similaridade densamente conectada para os quatro algoritmos avaliados. Barras denotam intervalo entre máximo e mínimo observados da quantia em questão dentro dos *splits* utilizados.

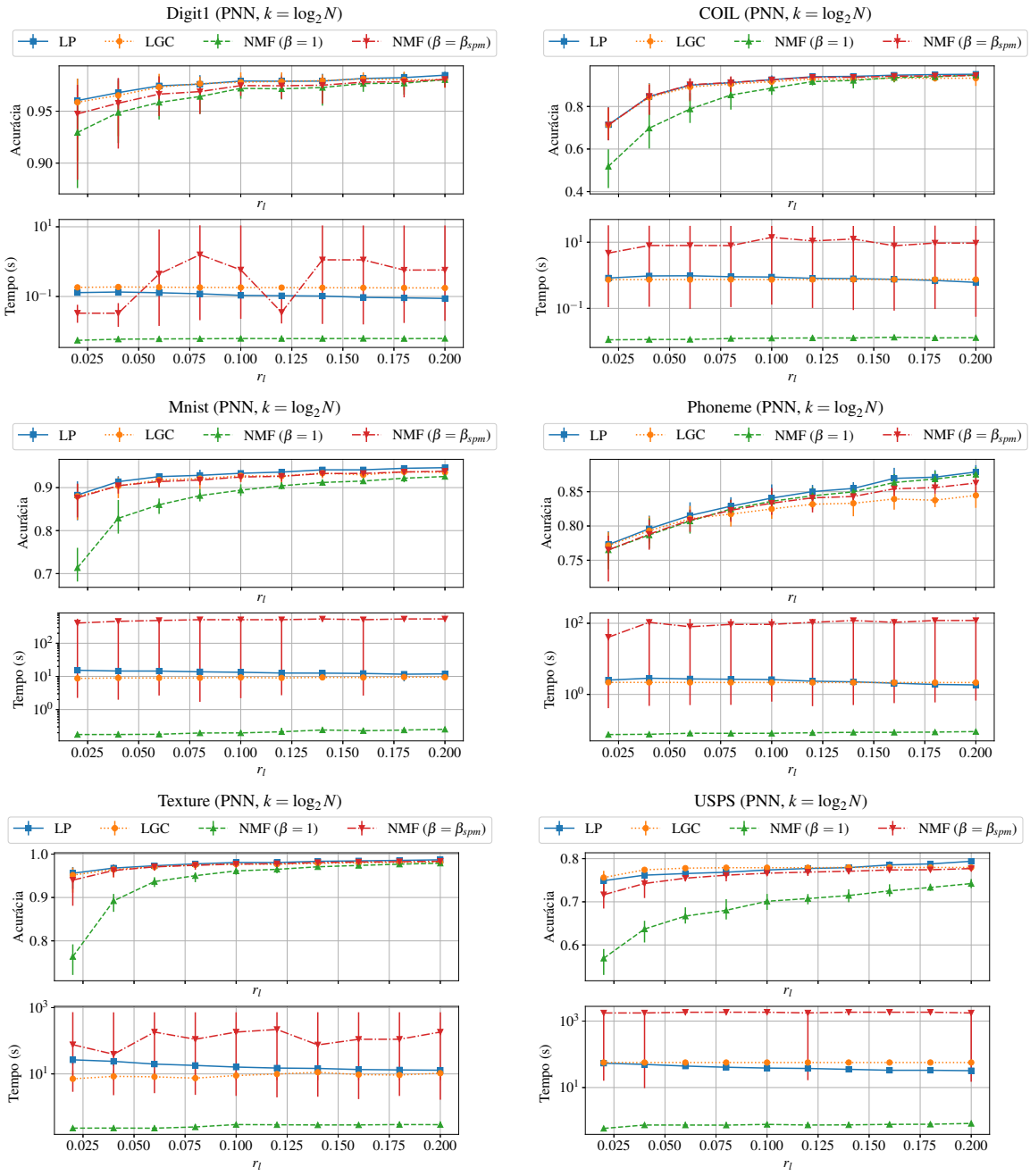

Figura 7 – Resultados da acurácia e tempo de execução em função de r_l para conjuntos de alta dimensão utilizando a similaridade PNN.

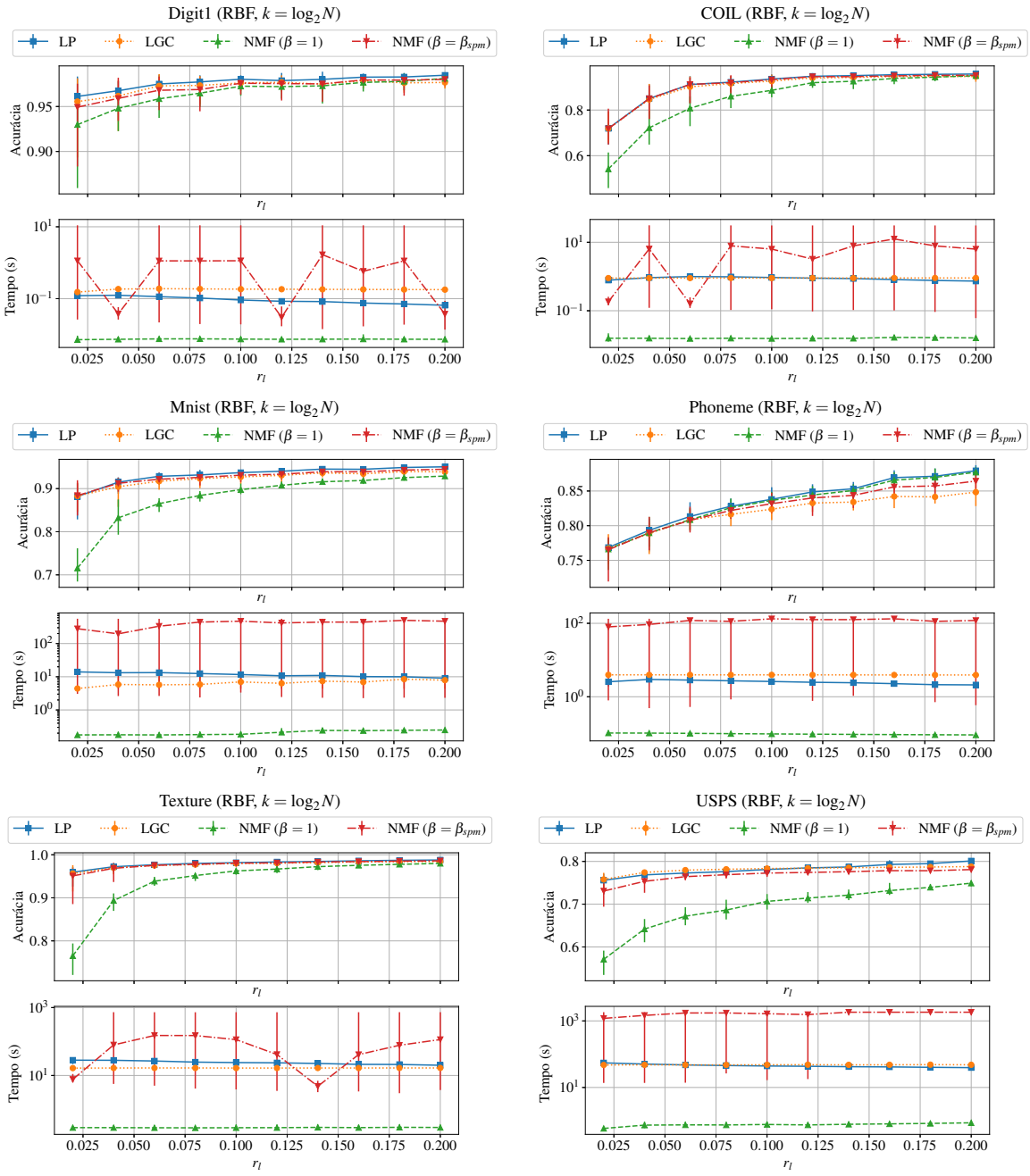

Figura 8 – Resultados da acurácia e tempo de execução em função de r_l para conjuntos de alta dimensão utilizando a similaridade RBF.

6 Conclusão

O presente trabalho focou-se em explorar as possibilidades de aprendizado semi-supervisionado com uma aproximação de campo médio do modelo Potts. Amparado por trabalhos anteriores foi possível inserir o modelo entre outros na literatura de SSL e realizar uma avaliação comparativa com algoritmos clássicos, além de elucidar seu funcionamento em diversas construções do grafo de similaridade G_D.

Em particular, $\beta = 1$ mostrou-se uma abordagem interessante frente à sua estabilidade em termos da construção de G_D, sendo capaz de operar em diversos regimes do parâmetro k. O ponto fraco desta abordagem são as situações em que r_l é muito inferior a $0,10$, que apresentaram uma acurácia significativamente inferior à dos demais métodos.

A utilização de $\beta = \beta_{spm}$ em conjuntos de alta dimensão mostrou que o mesmo tem um desempenho proibitivo em função do desaleramente crítico. Frente a essa dificuldade e às demais vantagens de $\beta = 1$, nas configurações estudadas este procedimento de ajuste mostrou-se o mais robusto para NMF. Como a possibilidade de um procedimento de ajuste para β dependente de r_l ainda é amparada pelos resultados deste trabalho, no futuro visa-se explorar esta hipótese mais a fundo.

Frente às limitações presentes de NMF, o principal desafio mostra-se o balanço entre acurácia e tempo de execução. Apesar do desaceleramento crítico ser um fenômeno próprio do modelo Potts, resta ainda averiguar se outras aproximações para sua distribuição de probabilidade oferecem vantagens sobre NMF.

Além disso, é de interesse também inserir o modelo no estudo de problemas mais localizados, nos quais a similaridades podem ser construídas de forma ainda mais precisa, como é o caso do processamento da linguagem natural. Também almeja-se realizar avaliações em dados com uma estrutura intrínseca, como é o caso da detecção de comunidades e da segmentação de imagens.

Referências

ALCALÁ-FDEZ, J. et al. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, Citeseer, v. 17, 2011. Citado na página 29.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2004. Citado 2 vezes nas páginas 13 e 14.

BEHNEL, S. et al. Cython: The best of both worlds. *Computing in Science & Engineering*, IEEE, v. 13, n. 2, p. 31–39, 2010. Citado na página 30.

BERTSEKAS, D. P.; HAGER, W.; MANGASARIAN, O. Nonlinear programming. In: _____. [S.l.]: Athena Scientific Belmont, MA, 1998. cap. 3.4. Citado na página 32.

BLATT, M.; WISEMAN, S.; DOMANY, E. Superparamagnetic clustering of data. *Physical review letters*, APS, v. 76, n. 18, p. 3251, 1996. Citado 5 vezes nas páginas 11, 22, 23, 25 e 26.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. *Semi-supervised learning*. [S.l.]: MIT Press, 2006. Citado 4 vezes nas páginas 14, 15, 16 e 29.

CHONG, Y. et al. Graph-based semi-supervised learning: A review. *Neurocomputing*, Elsevier, v. 408, p. 216–230, 2020. Citado 4 vezes nas páginas 10, 16, 17 e 33.

DAGUM, L.; MENON, R. Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engineering*, IEEE, v. 5, n. 1, p. 46–55, 1998. Citado na página 30.

ENGELEN, J. E. V.; HOOS, H. H. A survey on semi-supervised learning. *Machine Learning*, Springer, v. 109, n. 2, p. 373–440, 2020. Citado 9 vezes nas páginas 10, 14, 15, 16, 19, 20, 21, 24 e 33.

GABRIé, M.; TRAMEL, E. W.; KRZAKALA, F. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2015. p. 640–648. Citado na página 10.

GEORGES, A.; YEDIDIA, J. S. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, IOP Publishing, v. 24, n. 9, p. 2173, 1991. Citado na página 25.

GETZ, G.; SHENTAL, N.; DOMANY, E. Semi-supervised learning–a statistical physics approach. In: CITESEER. *In "Learning with Partially Classified Training Data"–ICML05 workshop*. [S.l.], 2005. Citado 3 vezes nas páginas 11, 23 e 25.

HULL, J. J. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 16, n. 5, p. 550–554, 1994. Citado na página 29.

KARASUYAMA, M.; MAMITSUKA, H. Adaptive edge weighting for graph-based learning algorithms. *Machine Learning*, Springer, v. 106, n. 2, p. 307–335, 2017. Citado na página 16.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 29.

LI, J.; WANG, F. Semi-supervised learning via mean field methods. *Neurocomputing*, Elsevier, v. 177, p. 385–393, 2016. Citado 4 vezes nas páginas 11, 22, 23 e 24.

MA, J.; XIAO, B.; DENG, C. Graph based semi-supervised classification with probabilistic nearest neighbors. *Pattern Recognition Letters*, Elsevier, v. 133, p. 94–101, 2020. Citado 3 vezes nas páginas 16, 19 e 20.

MARR, B. How much data do we create every day? the mind-blowing stats everyone should read. Forbes, 2018. Disponível em: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=2b84c53b60ba>. Citado na página 10.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-Hill, 1997. Citado na página 13.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of machine learning*. [S.l.]: MIT press, 2018. Citado 2 vezes nas páginas 13 e 14.

NIE, F.; WANG, X.; HUANG, H. Clustering and projected clustering with adaptive neighbors. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2014. p. 977–986. Citado na página 19.

OPPER, M.; SAAD, D. *Advanced mean field methods: Theory and practice*. [S.l.]: MIT press, 2001. Citado 2 vezes nas páginas 24 e 25.

OTT, T.; STOOP, R. Benefits and pitfalls of belief-propagation-mediated superparamagnetic clustering. *Physical Review E*, APS, v. 74, n. 4, p. 042103, 2006. Citado 4 vezes nas páginas 11, 23, 25 e 31.

PEARL, J. *Reverend Bayes on inference engines: A distributed hierarchical approach*. [S.l.]: Cognitive Systems Laboratory, School of Engineering and Applied Science, 1982. Citado na página 23.

RASCHKA, S. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, The Open Journal, v. 3, n. 24, abr. 2018. Disponível em: <http://joss.theoj.org/papers/10.21105/joss.00638>. Citado na página 29.

RITCHIE, D. M.; KERNIGHAN, B. W.; LESK, M. E. *The C programming language*. [S.l.]: Prentice Hall Englewood Cliffs, 1988. Citado na página 30.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 30.

SHI, C.; LIU, Y.; ZHANG, P. Weighted community detection and data clustering using message passing. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2018, n. 3, p. 033405, 2018. Citado 4 vezes nas páginas 11, 22, 23 e 25.

SOUSA, C. A. R. de; REZENDE, S. O.; BATISTA, G. E. Influence of graph construction on semi-supervised learning. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2013. p. 160–175. Citado 5 vezes nas páginas 17, 19, 20, 26 e 27.

SUBRAMANYA, A.; TALUKDAR, P. P. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Morgan & Claypool Publishers, v. 8, n. 4, p. 1–125, 2014. Citado na página 10.

WANG, F. et al. Semi-supervised mean fields. In: *Artificial Intelligence and Statistics*. [S.l.: s.n.], 2007. p. 596–603. Citado 4 vezes nas páginas 11, 22, 23 e 24.

WU, F.-Y. The potts model. *Reviews of modern physics*, APS, v. 54, n. 1, p. 235, 1982. Citado 2 vezes nas páginas 10 e 21.

YANG, H.; AHUJA, N. Clustering as physically inspired energy minimization. *Pattern Recognition*, Elsevier, v. 86, p. 265–280, 2019. Citado na página 10.

YEDIDIA, J. An idiosyncratic journey beyond mean field theory. *Advanced mean field methods: Theory and practice*, MIT Press, Cambridge MA, p. 21–36, 2001. Citado na página 24.

YEDIDIA, J. S.; FREEMAN, W.; WEISS, Y. Generalized belief propagation. *Advances in neural information processing systems*, v. 13, p. 689–695, 2000. Citado na página 23.

ZHOU, D. et al. Learning with local and global consistency. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2004. p. 321–328. Citado 4 vezes nas páginas 11, 16, 17 e 29.

ZHU, X.; GHAHRAMANI, Z. Learning from labeled and unlabeled data with label propagation. Citeseer, 2002. Citado 3 vezes nas páginas 11, 16 e 26.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado na página 14.